Blind Image Quality Assessment via Vector Regression and Object Oriented Pooling

Jie Gu, Gaofeng Meng, *Senior Member, IEEE*, Judith A. Redi, Shiming Xiang, *Member, IEEE*, and Chunhong Pan, *Member, IEEE*

***Abstract*—This paper presents an effective method based on vector regression and object oriented pooling for blind image quality assessment. Unlike previous models that map the extracted features directly to a quality score, the proposed vector regression framework yields a vector of belief scores for the input image. We explore the uncertainty factors in quality assessment and design the belief scores to measure the confidences of an image to be assigned to the corresponding quality grades. Moreover, we propose an object oriented pooling strategy to further improve the performance by incorporating semantic information of image contents. According to this strategy, regions occupied by objects will be assigned more weights in the pooling phase, leading to a more accurate quality assessment. Extensive experiments on benchmark datasets demonstrate that our approach achieves state-of-the-art performance and shows a great generalization ability.**

***Index Terms*—Convolutional neural network, image quality assessment, perceptual image quality, object oriented pooling, vector regression.**

I. Introduction

QUANTITATIVE evaluation on image and video quality is an important issue for many applications, such as image acquisition, restoration, compression, transmission, and enhancement. In most situations, subjective assessment is a natural way to evaluate the visual quality of images. However, subjective assessment is time-consuming, expensive, and cannot be used in the scenarios where a real-time and automated assessment is needed. Therefore, objective image quality assessment (IQA) [1] has gained growing attention in recent years.

According to the availability of reference images for quality assessment, existing objective IQA methods can be classified into three categories: full-reference (FR), reduced-reference (RR) and no-reference/blind (NR/B) methods. FR-IQA and RR-IQA methods evaluate the image quality by accessing either the entire (FR, [2]–[8]) or partial (RR, [9]–[11]) information about the reference image. In contrast, blind IQA (BIQA) methods are developed for the situations where the reference image is unavailable [12]–[16]. Particularly, general-purpose (*i.e.*, non-distortion-specific) BIQA methods do not limit themselves to specific types of distortions, and thereby are more widely used in real-world applications.

The current general-purpose BIQA methods can be roughly divided into two categories. One is the opinion-free models that do not require subjective scores for training. For example, learning a model without human opinion scores has been explored in [17] and [18]. Saha *et al.* [19] proposed a totally training-free model based on the scale invariance of natural images. The other category is the opinion-aware BIQA methods that are usually developed within a single regression framework. The main idea is to learn a regression model that maps the extracted features directly to a quality score [20].

Natural scene statistics (NSS) features are the most popular features adopted by BIQA methods [12]–[14], [21]–[24]. Blind image quality index (BIQI) [21] is one of the pioneering models, which uses distorted image statistics to build a two-step framework for BIQA. In [22], a group of low-level features, derived from natural image statistics, texture features and blur/noise estimation, are used to build the LBIQ metric. Wavelet and DCT transform coefficients are explored in DIIVINE [12] and BLIINDS-II [13], respectively. Mittal *et al.* presented BRISQUE [14] in the spatial domain with the mean subtracted contrast normalized (MSCN) coefficients. In [23], NSS features are combined with free energy principle based features to build the NFERM metric. Zhang *et al.* [25] learned a multivariate Gaussian model based on the NSS features derived from multiple cues. Ghadiyaram and Bovik [24] extracted a rich set of statistical features to achieve good quality prediction on authentically distorted images.

Instead of designing hand-crafted features, feature learning methods aim to acquire quality-aware representations from raw images. Unsupervised feature learning has been explored in [15], [26]–[29]. Ye *et al.* [15] proposed to encode patches via soft-assignment and max pooling to obtain general features. Semantic obviousness metric [28] combines two types of

Manuscript received August 13, 2017; revised September 15, 2017; accepted September 24, 2017. Date of publication October 11, 2017; date of current version April 17, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61370039, Grant 91646207, and Grant 91338202, and in part by the Beijing Nature Science Foundation under Grant 4162064. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wolfgang Hürst. *(Corresponding author: Jie Gu.)*

J. Gu and S. Xiang are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: jie.gu@nlpr.ia.ac.cn; smxiang@nlpr.ia.ac.cn).

G. Meng and C. Pan are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: gfmeng@nlpr.ia.ac.cn; chpan@nlpr.ia.ac.cn).

J. A. Redi is with the Department of Intelligent Systems, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: J.A.Redi@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2017.2761993

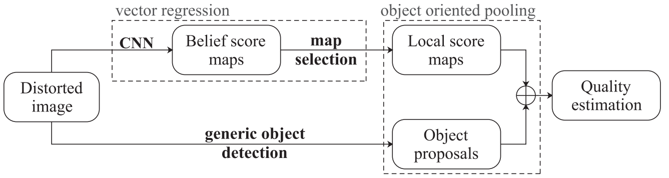


Fig. 1. Pipeline of the proposed method. It consists of two modules: vector regression and object oriented pooling.

features. One focuses on the local characteristics and the other measures the semantic obviousness of an image. Supervised feature learning has also been used in many BIQA methods, where the feature extractor and regression model are learned jointly. Recently, deep learning has made a great progress in a variety of vision tasks [30]–[33]. Researchers have shown an increasing interest in applying deep neural networks to BIQA [16], [34]–[39].

During the training of existing BIQA models, the mean opinion scores (MOSs) are given as the ground-truth quality scores, which are obtained through subjective experiments. We notice that, in subjective tests, different subjects may assign different quality scores to the same image due to many factors, such as physiological differences, personal preferences, and individual differences. This observation indicates that the uncertainty due to the individual differences could be well explored to improve the performance of BIQA.

In this paper, we propose an effective approach based on vector regression and object oriented pooling (VROP) for BIQA. Fig. 1 shows the pipeline of our method. At the stage of vector regression, we introduce multiple quality grades and use a belief score vector to measure the confidences of an input image to be assigned to these grades. In practice, the vector regression is implemented by a convolutional neural network (CNN). The belief score maps generated by the CNN are further transformed into local score maps for global score pooling. At the stage of score pooling, we propose an object oriented pooling strategy to convert the local score maps into image level quality score. Since objects are more likely to be noticed as shown in some previous studies [40], regions occupied by objects will be more weighted in the proposed pooling strategy.

The contributions of our work are three-folds.

1) A vector regression framework is proposed for BIQA. In this framework, we explore the uncertainty in quality assessment and introduce a vector of belief scores to measure the probabilities of an image to be assigned to the corresponding quality grades. Experimental results show that this framework can achieve better performance over the state-of-the-art methods.
2) The proposed vector regression framework is an open framework. Experimental results on authentically distorted images indicate that it can be integrated with different CNNs and benefits the performance regardless of the network architecture.
3) An object oriented pooling strategy is proposed to further improve the performance at the stage of the global score estimation. This strategy assigns more weights to object-like regions within which patches are more likely to be object proposals.

The remainder of this paper is organized as follows. Section II briefly reviews the related works. Section III introduces the basic principle of our approach. Section IV describes the proposed BIQA method in detail. We report the experimental results and present the discussions in Section V and VI respectively. Finally, Section VII concludes the work.

II. Related Work

In this section, we introduce previous related works, including a brief review of neural network based BIQA methods, representation for multi-level image quality and generic object detection.

A. Neural Networks for BIQA

Deep learning has promoted the developments of many visual tasks, including image classification [30], [41], object recognition [31], [42], semantic segmentation [32], [43], visual tracking [33], [44], etc. Tremendous progress has been made due to the representative power of deep neural network. Recently, researchers show increasing interests in applying neural networks to BIQA.

Deep belief network (DBN) has been explored to extract general quality features in some previous works. An early version of FRIQUEE [34] combines a DBN with an SVM, where the network is used to generate more complex representations from pre-extracted features. Tang *et al.* [35] presented a BIQA method using a rectifier neural network. In their work, a DBN is designed to provide complex feature representations and finally a Gaussian process is used to obtain the image quality score. Hou *et al.* [36], [37] treated BIQA as a classification problem. They first classify the features into five quality grades with a DBN, and then the qualitative labels are converted into a numerical score via quality pooling.

CNN was first applied to BIQA in [16]. Kang *et al.* used a shallow network to estimate the quality scores of small non-overlapping patches, and then the predicted scores are averaged to obtain the image level quality score. Lu *et al.* [38] proposed a CNN-based multi-patch aggregation architecture, which conducts the feature extraction and aggregation function learning jointly.

In this paper, the key idea is to use CNN to estimate the MOS of an input image by first estimating a belief score vector. The predicted belief scores indicate the probabilities of an input image being assigned to the corresponding quality grades. This approach is radically different from the previous ones, as the MOS is expressed as a membership function to multiple grades in our algorithm.

B. Representation for Multi-Level Image Quality

Multi-level quality representation has been explored in some previous BIQA works [17], [26], [27], [29]. They are typically based on unsupervised feature learning, and the main idea is to use some pre-extracted features to encode images. The specific encoding strategy varies, but eventually a vector representation is learned by measuring the similarities between the pre-extracted features and the features of the input image. In general, the pre-extracted features cover multiple quality levels, thus the generated representation can be considered as a multi-level quality representation.

Specifically, Ye and Doermann [26] proposed to encode images with a codebook consisting of Gabor-filter-based local features. An image is presented as a histogram of occurrence counts for different codewords. He *et al.* proposed SRNSS [27] based on sparse representation. In their method, the dictionary is constructed by combining the NSS features and MOSs of training images. QAC [17] divides the quality scale into multiple levels. In this work, the clustering centroids at different quality levels, served as a codebook, are used to encode images. Wu *et al.* [29] extracted both the frequency and spatial-frequency features from all three YCbCr channels. A label transfer model is then developed to estimate the quality.

Our belief score vector is somewhat similar to a multi-level representation of image quality score. The difference lies in that the proposed method is implemented in a supervised manner and does not require a large set of features (*e.g.*, codewords) to explicitly generate a quality-aware representation. It benefits from supervised feature learning and is free from pre-designed encoders.

C. Generic Object Detection

Generally, humans can spontaneously perceive objects even before recognizing them. Inspired by that, a variety of recent works concentrate on designing object detection methods, aiming at producing a set of class-independent object proposals [45]–[49]. Typically, it can be applied to reduce the search space of local regions, which may be helpful to some visual tasks like object detection [31], [42]. Existing generic object detection methods can be broadly grouped into two classes, namely the ones generating image windows and the ones producing segmented object hypotheses.

Among the former, by combining several complementary cues in a Bayesian framework, Alexe *et al.* [45] designed an objectness measure to distinguish the object windows from the background windows. In [46], Uijlings *et al.* proposed a data-driven selective search strategy, which is subject to the considerations about scale and diversification. Cheng *et al.* [47] proposed a simple yet powerful feature for estimating the objectness of image window, called 'BING', which is extremely efficient (300 fps on a single CPU) as the computation requires only several atomic CPU operations.

On the contrary, other works focus on pixel-accurate proposals that encode informative boundary shape cues. Arbeláez *et al.* [48] proposed a unified approach for both hierarchical segmentation and proposal extraction, called multiscale combinatorial grouping (MCG). The cores of MCG are a fast algorithm for normalized-cut segmentation and an efficient grouping strategy for combining multiscale regions. Krähenbü and Koltun [49] presented the geodesic object proposal (GOP) method, which is substantially fast and outperforms the state of the art especially in object shape accuracy. These two advantages of GOP exactly correspond with our demands for the object oriented pooling strategy.

III. Basic Principle of Vector Regression

During the assessment of perceptual quality of an image, different subjects may assign different quality scores to the given image because of personal preferences and physiological differences. To model this kind of uncertainty, we introduce a probability distribution $P(Q|\mathbf{x})$ to describe quality as perceived by different people in a population, where $\mathbf{x}$ represents an image and Q is a continuous variable that indicates the estimated quality score of this image.

Discretization of the quality continuum is quite common in literature related to subjective quality assessment. Likert or categorical scales (such as the widely used absolute category rating scale) are most often used to collect quantitative judgments of image quality. Similarly, we divide the numerical scoring scale into several ordered intervals. Denote the center of the k-th interval as μ_k, we have:

$$\mu_1 < \mu_2 < \ldots < \mu_K \ , \tag{1}$$

where K is the number of the ordered intervals. Basically, these intervals correspond to K quality grades, and the probability of the input image $\mathbf{x}$ belonging to the k-th quality interval, denoted by P_k, can be easily computed from $P(Q|\mathbf{x})$.

In practice, however, to directly estimate the probability P_k is difficult as the number of subjective evaluations in benchmark dataset is insufficient to accurately estimate the distribution. In this study, we assume the distribution $P(Q|\mathbf{x})$ is symmetric and unimodal, and divide the scoring scale into equally sized intervals. A similar assumption, *e.g.*, Gaussian distributions for image quality ratings, was also adopted in [50]. Under this assumption, the probability P_k $(k = 1, 2, \ldots, K)$ is positively related to the distance from the mean value of the distribution (can be estimated from the MOS) to the center of the k-th interval. That is to say, the smaller the distance is, the larger the probability will be. Therefore, we can use an implicit way to measure the probability.

We introduce a belief score vector to describe the probabilities of an image being assigned to different quality grades. The belief score of the k-th quality grade is defined as the distance from the MOS of an image to the center of the corresponding interval, *i.e.*,

$$s_k = y - \mu_k \quad (k = 1, 2, \ldots, K) \ , \tag{2}$$

where s_k $(k = 1, 2, \ldots, K)$ is the defined belief score and y is the MOS of an image. Generally, the smaller the value of $|s_k|$ is, the larger P_k will be. That means the image is more likely to be assigned to the k-th quality grade by the population.

Following the ideas above, we propose a vector regression framework for BIQA. Previous BIQA methods generally learn a single regression model that maps the extracted features directly to a quality score. In contrast, the proposed vector regression framework uses a belief score vector to represent the quality score. Our model estimates the interrelated belief scores instead

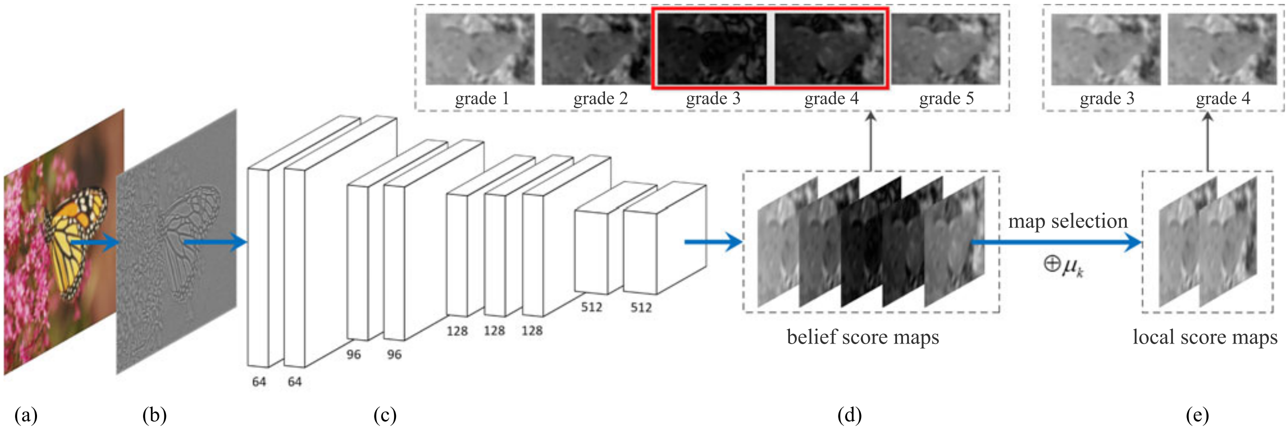


Fig. 2. Flowchart of the vector regression module. (a) Original image. (b) Locally normalized image. (c) Fully convolutional neural network. (d) Defined belief score maps and their corresponding quality grades. (e) Local score maps. In this case, the 3-th and 4-th grades are selected for global score pooling (the third and fourth belief score maps have the smallest absolute means, as shown in a red box). The operator $\oplus\mu_k$ represents element-wise addition with μ_k as in (7).

of a single MOS as in previous ones. In practice, the vector regression is implemented by a convolutional network. More details will be given in next section.

IV. BIQA via Vector Regression and Object Oriented Pooling

The proposed method VROP consists of two modules, *i.e.*, vector regression and object oriented pooling. The vector regression is implemented by a deep fully convolutional network. An object oriented pooling strategy is further used to convert the local score maps into a single objective score.

A. CNN-Based Vector Regression

The flowchart of the vector regression module is shown in Fig. 2. This module consists of three main steps, *i.e.*, image preprocessing, belief score mapping and score map selection. In this section, we will describe these steps and provide the details on module training.

1) Network and Preprocessing: The deep network is built by stacking 3×3 convolutional layers as shown in Fig. 2(c). There are 7 convolutional layers with 64,64,96,96,128,128, 128 channels respectively. The convolutional stride is set as 1 and the padding is also set as 1 to preserve the spatial resolution. Max pooling is applied after the second, fourth and seventh layers over a 3×3 window with stride 2. During training, the last pooling layer is followed by three fully connected layers: The first two have 512 channels each, the third one performs K-way regression and thus contains K channels ($K = 5$ in Fig. 2). In addition, the non-linearity mapping is implemented by rectified linear units (ReLUs) [51], and Dropout [52] is applied after the second fully connected layer as a regularizer to avoid overfitting.

We employ a local contrast normalization method as in [14], [16] to preprocess the input image. Unlike common contrast normalization, the local normalization computes the mean and variance for each pixel. Assuming the gray image is I, we compute the normalized image as

$$\hat{I}(i,j) = \frac{I(i,j) - u(i,j)}{\sqrt{\sigma(i,j)} + C}, \tag{3}$$

$$u = \omega \otimes I, \tag{4}$$

$$\sigma = \omega \otimes (I \odot I) - u \odot u\ , \tag{5}$$

where i, j are spatial indices, u and σ represent the local mean and variance, C is a positive constant to avoid instability when σ is close to zero, and ω is a Gaussian function. The operators $\otimes$ and $\odot$ represent convolution and element-wise product respectively. The size and sigma value of the Gaussian function are set as default values, *i.e.*, 3×3 and 1.5 respectively. C is set to be 1. Note that the normalized image exhibits homogeneous and uniform appearance [14], which benefits the subsequent training processes.

2) Generation of Training Data: To train the deep network with limited labeled data, we divide image into patches and feed them to the network during training. Specifically, the same data augmentation method as in [16] is adopted. We collect samples by cropping non-overlapping 32×32 patches from locally normalized images. For each patch, by assigning its MOS as the ground-truth score of the source image, the belief score vector can be computed by (2). The CNN is trained on these cropped image patches and the associated belief score vectors.

However, the above data augmentation strategy is inappropriate when the training images contain inhomogeneous impairments [16]. Therefore, in our study, we modify the data augmentation strategy by excluding those patches with quality scores different from their source images. To this end, in the training phase, we use a FR measure VSI [8] to generate a reference quality score for each cropped patch. A patch will be discarded if the difference between y and $\hat{y}$ is larger than a certain threshold δ, *i.e.*, $|y - \hat{y}| > \delta$, where y is the ground-truth score of its source image and $\hat{y}$ is the estimated reference score of the patch.

3) Loss Function: We use the squared error with weight decay as the loss function

$$L\left(\mathbf{W}, \mathbf{b}\right) = \frac{1}{N}\sum_{i=1}^{N} \|f_{\mathbf{W},\mathbf{b}}\left(\mathbf{p}_i\right) - \mathbf{s}_i\|_2^2 + \lambda\|\mathbf{W}\|_F^2 \ , \qquad (6)$$

where $\mathbf{p}_i$ and $\mathbf{s}_i$ represent the sampled patch and label vector of belief scores, $f_{\mathbf{W},\mathbf{b}}\left(\mathbf{p}_i\right)$ is the output vector of the network, $\mathbf{W}$ and $\mathbf{b}$ are the weights and biases of the network respectively, and λ is the hyper-parameter. Different from the classification-based models with the softmax loss as the learning objective, our approach is constructed under the regression framework in terms of the least square loss.

The loss function is optimized by stochastic gradient descent (SGD) with a momentum of 0.9 and a mini-batch size of 256. We initially set the learning rate as 0.01 and then decrease it by a factor of 0.1 at regular intervals. The iteration is performed for about 100 epochs and the learning rate is decreased about 3 times in total. Moreover, the weight decay is set as 0.0005, and the dropout ratio is set to be 0.5 in the training stage and divided by 2 in the testing stage.

4) Local Score Maps Generation: At testing stage, a fully convolutional network is built by transforming the fully connected layers into convolutional layers [43]. Naturally, we can feed a locally normalized image instead of an image patch to the adapted fully convolutional network and obtain multiple belief score maps, denoted by $\mathbf{S}_k$, $k = 1, 2, \ldots, K$. These maps can be considered as a set of belief score vectors estimated by the original network on some particular patches.

Basically, the smaller the absolute mean of $\mathbf{S}_k$ is, the more likely the quality of the input image can be described as the k-th quality grade. In our model, an object-based pooling strategy is designed to obtain the image quality score. Before that, two belief score maps with the smallest absolute means are picked out to compute two local score maps, *i.e.*,

$$\mathbf{L}_k = \mathbf{S}_k \oplus \mu_k \quad \left(k = p; p+1\right) \ , \qquad (7)$$

where $\mathbf{L}_k$ denotes the local score map, p and $p+1$ are the indices of the selected belief score maps, and the operator $\oplus$ represents element-wise addition. Note that the selected two maps are always adjacent to each other and that either $\mathbf{S}_p$ or $\mathbf{S}_{p+1}$ has the minimum absolute mean among the generated belief score maps.

The main purpose of retaining only two belief score maps is to reduce the redundancy of the scores. The basic principle of the map selection is illustrated in Fig. 2(d). It can be seen that the selected two quality grades are the most relevant ones to the ground-truth score of test image. That is to say, the given image has a higher probability of being assigned to these two grades by the population.

B. Object Oriented Pooling

Once the local score maps are predicted, directly averaging these two quality maps is an intuitive and efficient approach to obtain the image level quality score. Let p and $p+1$ be the indices of the selected belief score maps. Denote W and H the width and height of the generated local score maps. The global image quality score can be simply defined as the average over $\mathbf{L}_p$ and $\mathbf{L}_{p+1}$, *i.e.*,

$$S_{img} = \frac{1}{2WH}\sum_{i,j}\left(\mathbf{L}_p + \mathbf{L}_{p+1}\right) \ , \qquad (8)$$

where S_{img} is the global quality score, i and j indicate the location of $\mathbf{L}_p$ and $\mathbf{L}_{p+1}$.

In practice, we notice that the performance of our approach can be further improved by incorporating semantic information of image contents. The reason for the improvement may be due to the visual attention mechanism that humans always fixate some particular regions of an image [53], [54]. Some previous studies have shown that most of the time humans focus on object-like regions when looking at an image [40].

Following the ideas above, we propose a statistics-based object oriented pooling strategy. Our pooling strategy is similar to SOM [28] for using generic object detection methods to select object-like regions. The difference is that the local regions are selected to compute the global quality score directly rather than to extract local features as in SOM (which can be seen as an implicit way to assign weights). Specifically, a set of object proposals are first extracted by the geodesic object proposal (GOP) method [49]. The generated object candidates can be seen as an over-complete coverage of the object-like regions, retaining the detailed shape and boundary. To reduce the redundancy of these proposals, only a subset is preserved for pooling. The subset is selected randomly, and its size, denoted as M, is a parameter of our approach. After that, we map these selected proposals to the local score maps to compute their respective quality scores.

Denote r_m^1 and r_m^2 the quality scores of the m-th selected proposal obtained from $\mathbf{L}_p$ and $\mathbf{L}_{p+1}$, respectively. Let o_m be the mask of the m-th selected proposal. $o_m\left(i,j\right) = 1$ indicates that the pixel at the location $\left(i,j\right)$ belongs to this proposal, and $o_m\left(i,j\right) = 0$ otherwise. Then r_m^1 can be computed as (r_m^2 can be obtained similarly)

$$\begin{aligned} r_m^1 &= \sum_{i,j}\mathbf{L}_p \odot \mathbf{F} \\ \mathbf{F} &= \frac{o_m}{\sum\limits_{i,j} o_m\left(i,j\right)}, \end{aligned} \qquad (9)$$

where i and j indicate the location, and the operator $\odot$ represents element-wise product. Note that o_m in (9) has been resized to the same size as the local score maps. An illustration of the pooling process is shown in Fig. 3. Finally, the image level quality score can be computed as

$$S_{img} = \frac{1}{2M}\sum_{m}\left(r_m^1 + r_m^2\right) \ . \qquad (10)$$

V. Experiments

To evaluate the performance of our approach, we implemented three experiments on six widely used datasets. The experiments on two benchmark datasets [55], [56] aim to validate how the objective assessment corresponds to subjective evaluation. The cross-dataset evaluation tests the generalization ability of the proposed method.

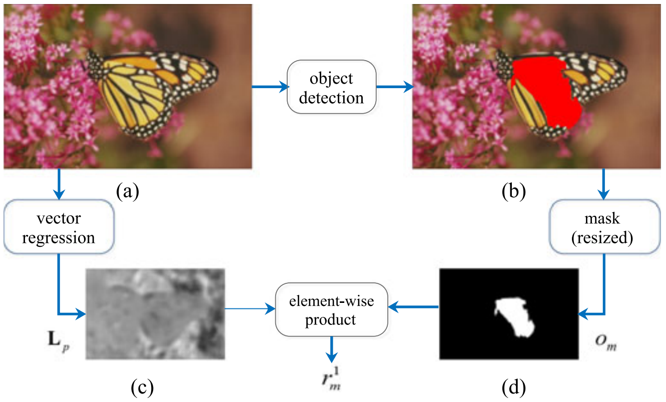


Fig. 3. Illustration of the object oriented pooling strategy. (a) Original image. (b) m-th selected proposal. (c) Local score map $\mathbf{L}_p$. (d) Mask o_m of the m-th selected proposal. $o_m(i,j) = 1$ indicates that the pixel at the location (i,j) belongs to this proposal, and $o_m(i,j) = 0$ otherwise. Note that o_m needs to be resized to the same size as $\mathbf{L}_p$ as shown in (d). r_m^1 is the predicted quality score of the m-th selected proposal obtained from $\mathbf{L}_p$.

TABLE I
BENCHMARK DATASETS FOR EVALUATING BIQA METHODS

Dataset	Source Images	Distorted Images	Distortion Types
LIVE	29	779	5
TID2008	25	1700	17
TID2013	25	3000	24
CSIQ	30	886	6
IVC	10	185	4
MICT/LCD	14	168	2

A. Experimental Protocol

Datasets: We conduct the experiments on six public IQA datasets, including LIVE [55], TID2008 [56], TID2013 [57], CSIQ [58], IVC [59] and MICT/LCD [60]. The characteristics of these six datasets, namely the number of source images, the number of degraded images and the number of distortion types, are summarized in Table I.

Evaluation: Two commonly used performance metrics are employed to evaluate the IQA methods, *i.e.*, Spearman rank order correlation coefficient (SROCC) and linear correlation coefficient (LCC). These two metrics are used to measure the monotonicity and the linear dependence between the predicted objective scores and the subjective scores, respectively. As the range of the objective scores may be different from that of the subjective scores, a nonlinear mapping needs to be performed before computing the LCC metric. Generally, the mapping function is a logistic function as suggested by [55]

$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp\left(\beta_2 \left(x - \beta_3\right)\right)} \right) + \beta_4 x + \beta_5 , \tag{11}$$

where β_i, $i = 1, 2, \ldots, 5$, are the parameters to be fitted.

In our experiments, we also compare the object oriented pooling strategy with the average pooling strategy. The two methods are abbreviated to VRAP and VROP (vector regression and average/object pooling), respectively. Note that we only report results on distorted images in the experiments.

TABLE II
PERFORMANCE COMPARISON ON THE SPECIFIC DISTORTION TYPES OF LIVE AND THE WHOLE LIVE DATASET

SROCC	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.870	0.885	0.942	0.763	0.874	0.866
SSIM [2]	0.939	0.946	0.964	0.907	0.941	0.913
FSIM [5]	0.970	0.981	0.967	0.972	0.949	0.964
DIIVINE [12]	0.913	0.910	0.984	0.921	0.863	0.916
BLIINDS-II [13]	0.929	0.942	0.969	0.923	0.889	0.931
BRISQUE [14]	0.914	0.965	0.979	0.951	0.877	0.940
CORNIA [15]	0.943	0.955	0.976	**0.969**	0.906	0.942
CNN [16]	0.952	**0.977**	0.978	0.962	0.908	0.956
NFERM [23]	0.942	0.965	0.984	0.922	0.863	0.941
DLIQA-R [36]	0.933	0.914	0.968	0.947	0.857	0.929
SOM [28]	0.947	0.952	0.984	**0.976**	**0.937**	**0.964**
VRAP	**0.953**	**0.979**	**0.988**	0.924	0.934	0.951
VROP	**0.963**	0.976	**0.984**	0.956	**0.939**	**0.967**
LCC	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.873	0.876	0.926	0.779	0.870	0.856
SSIM [2]	0.921	0.955	0.982	0.893	0.939	0.906
FSIM [5]	0.910	0.985	0.976	0.978	0.912	0.960
DIIVINE [12]	0.922	0.921	0.988	0.923	0.888	0.917
BLIINDS-II [13]	0.935	0.968	0.980	0.938	0.896	0.930
BRISQUE [14]	0.923	0.973	0.985	0.951	0.903	0.942
CORNIA [15]	0.951	0.965	0.987	**0.968**	0.917	0.935
CNN [16]	0.953	0.981	0.984	0.953	0.933	0.953
NFERM [23]	0.955	0.982	**0.992**	0.937	0.888	0.946
DLIQA-R [36]	0.953	0.948	0.961	0.950	0.892	0.934
SOM [28]	0.952	0.961	**0.991**	**0.974**	**0.954**	**0.962**
VRAP	**0.963**	**0.984**	0.983	0.940	0.938	0.953
VROP	**0.975**	**0.984**	0.983	0.959	**0.951**	**0.968**

Parameter setting: The threshold δ, for excluding training patches, is set as 15 percent of the data range in the training dataset (*e.g.*, differential mean opinion score (DMOS) in LIVE is in the range $[0, 100]$, thus δ is set as 15). During testing, a number of object proposals are randomly selected for pooling. This number, *i.e.*, M, is set to be one-third of the total number of the generated object proposals. Moreover, the number of the intervals K is set as 5, and the centers of the intervals, *i.e.*, $(\mu_1, \mu_2, \ldots, \mu_5)$, are set to be equidistant, *e.g.*, $(0, 25, 50, 75, 100)$ for the LIVE dataset and $(0, 2, 4, 6, 8)$ for the TID dataset.

B. Evaluation on LIVE

To ensure a fair comparison, the same experimental strategy as in [14]–[16], [23], [28] is adopted. We group the distorted images in LIVE according to their reference images, and randomly select 23 groups for training, while retaining the other 6 groups for testing. To remove the influence of random selection, the results are all reported as the medians of 20 train-test iterations. Table II shows the experimental results, where the best two BIQA methods are highlighted in boldface. Eight representative BIQA methods and three FR-IQA methods are tested for comparison. The results of the BIQA methods, including DIIVINE [12], BLIINDS-II [13], BRISQUE [14], CORNIA [15], CNN [16], NFERM [23], DLIQA-R [36] and SOM [28], are taken from the original papers.

TABLE III
PERFORMANCE COMPARISON ON THE TID2008 DATASET

	SSIM [2]	FSIM [5]	BRISQUE [14]	CORNIA [15]	CNN [16]	NFERM [23]	SOM [28]	VRAP	VROP
LCC	0.857	0.913	0.795	0.837	0.873	0.849	0.846	0.893	**0.913**
SROCC	0.878	0.926	0.768	0.813	0.862	0.842	0.808	0.900	**0.911**

(The results of CORNIA and CNN are reported by original authors, and SOM is implemented by ourselves.)

We conduct both distortion-specific and non-distortion-specific experiments to evaluate the proposed approach. The purpose of the distortion-specific experiments is to evaluate the performance when there are only images with one particular type of distortion. To this end, we split the training and testing parts only on specific distortion types. As for the non-distortion-specific experiments, the whole dataset is randomly split into the training and testing parts in each iteration. The results of the non-distortion-specific experiments are listed in the last column in Table II.

It can be seen that the proposed method works consistently well in both distortion-specific and non-distortion-specific experiments. Specifically, our approach obtains convincing results on each of the five distortions, especially on JPEG2000 compression (JP2K), JPEG compression (JPEG), Gaussian noise (WN) and fast fading (FF). As for the overall evaluation, both VRAP and VROP achieve state-of-the-art results compared with other BIQA and FR-IQA methods. Our method achieves a comparable performance with SOM, which is a no-reference method and has the best performance on the LIVE dataset. SOM combines local features with semantic obviousness features for a better performance. In contrast, the good performance of our approach is owing to the proposed vector regression framework and object oriented pooling strategy. The former explores the uncertainty in quality assessment and the latter focuses on object-like regions for a more accurate prediction.

C. Evaluation on TID2008

To further examine the performance of our approach, more challenging evaluations are performed in this section. As previously done in [15], we conduct non-distortion-specific experiments on the first thirteen distortions in the TID2008 dataset. The other four distortions, namely non eccentricity pattern noise (NEPN), local block-wise distortions of different intensity (LBD), intensity shift (IS) and contrast change (CC), are not included as they are either very inhomogeneous or highly subjective for BIQA [15].

Similarly, the degraded images in TID2008 are divided into 25 groups according to their reference images. The results are reported as the median values across 20 train-test iterations, where 20 groups are randomly selected for training and the remaining 5 groups for testing.

In Table III, two FR-IQA methods, SSIM [2] and FSIM [5], are used as the baseline. Five representative BIQA methods, *i.e.*, BRISQUE [14], CORNIA [15], CNN [16], NFERM [23] and SOM [28], are tested for comparison. One can see from Table III that the proposed method outperforms the other five BIQA methods and approaches the FR-IQA measure FSIM, given only distorted images. Furthermore, as we can see in Table II and Table III, the performance decreases to some extent if the object oriented pooling is not incorporated, *e.g.*, in non-distortion-specific experiments, LCC decreases 1.5% on LIVE and 2.0% on TID2008. More experiments will be performed to demonstrate the effectiveness of the object oriented pooling strategy in the next section.

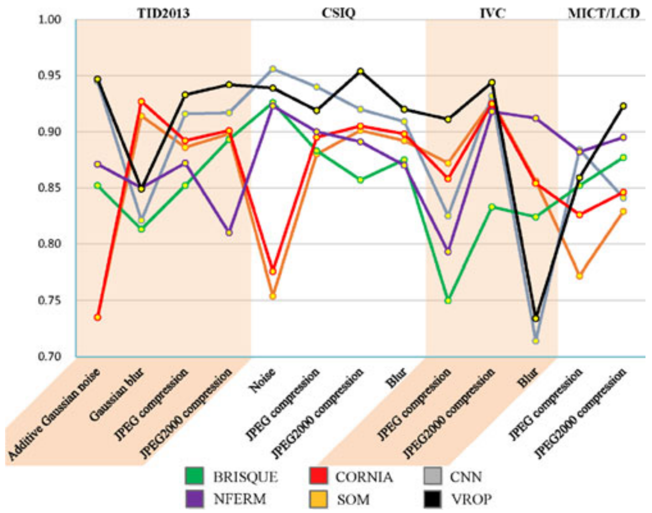


Fig. 4. SROCC metrics of different methods on group 1, where the types of distortions have examples in LIVE. The vertical and horizontal axes represent the SROCC metric and the distortion types in each dataset respectively.

D. Cross Dataset Evaluation

In this section, we demonstrate that the performance of our approach is independent of testing datasets. We train our model on the entire LIVE dataset and then test the performance on the other four datasets: TID2013, CSIQ, IVC and MICT/LCD. Since many distortion types in the testing datasets do not appear in LIVE, we separate the test images into two groups. Group 1 contains those types of distortions appearing in LIVE and group 2 contains the rest. We perform the nonlinear mapping on the predicted objective scores and compute the evaluation criteria as the way of evaluating FR measures. Five representative BIQA models are included for comparison. The source codes of these compared models are obtained from the original authors except CNN and SOM, which are implemented by ourselves.

Figs. 4 and 5 show the SROCC metrics of BRISQUE, CORNIA, CNN, NFERM, SOM and our VROP on the two groups, respectively. It can be seen from Fig. 4 that VROP outperforms the other five models on most types of distortions in group 1. Specifically, it reaches over 0.9 for 10 of 13 distortion types,

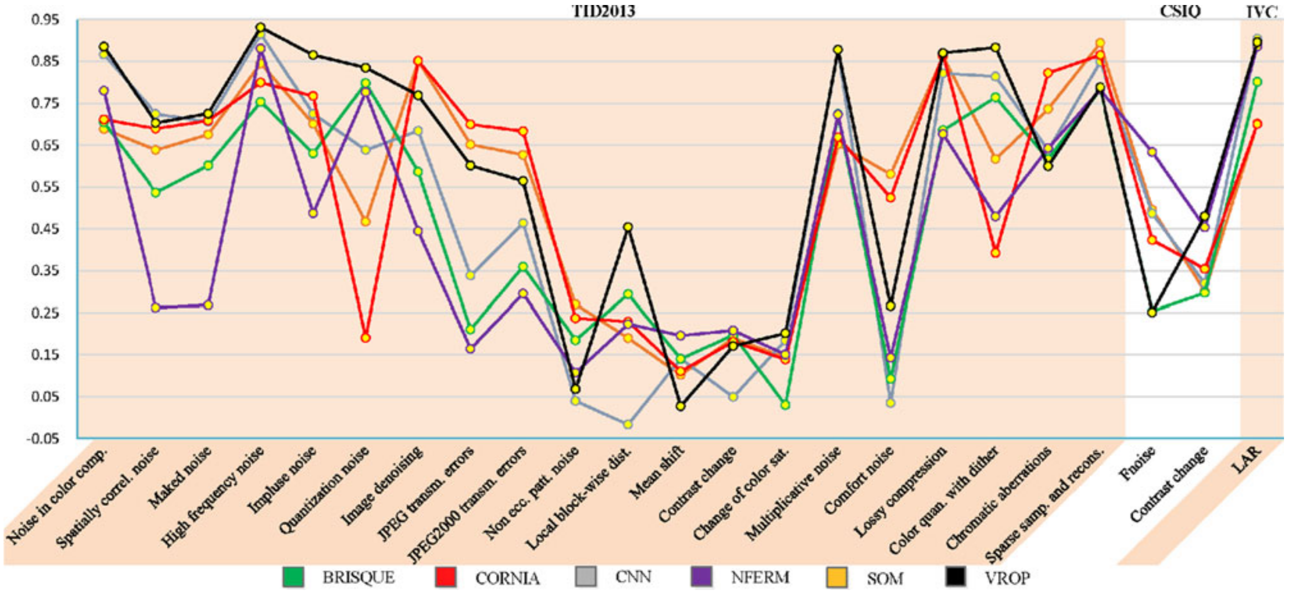


Fig. 5. SROCC metrics of different methods on group 2, where the types of distortions do not appear in LIVE. The vertical and horizontal axes represent the SROCC metric and the distortion types in each dataset respectively.

TABLE IV
THE RESULTS OF NON-DISTORTION-SPECIFIC EXPERIMENTS ON FOUR TESTING DATASETS

SROCC	CSIQ	TID2013	IVC	MICT/LCD
BRISQUE [14]	0.882	0.842	0.768	0.863
CORNIA [15]	0.889	0.879	0.889	0.832
CNN [16]	0.927	0.892	0.844	0.846
NFERM [23]	0.897	0.869	0.813	0.880
SOM [28]	0.880	0.873	0.887	0.802
VRAP	0.925	0.905	0.893	0.858
VROP	**0.933**	**0.923**	**0.903**	**0.896**
LCC	CSIQ	TID2013	IVC	MICT/LCD
BRISQUE [14]	0.898	0.863	0.768	0.861
CORNIA [15]	0.903	0.888	0.883	0.821
CNN [16]	0.934	0.890	0.839	0.842
NFERM [23]	0.914	0.875	0.816	0.881
SOM [28]	0.902	0.894	0.879	0.791
VRAP	0.936	0.911	0.889	0.861
VROP	**0.945**	**0.927**	**0.901**	**0.898**

The models are trained on the LIVE dataset and then tested on the images with all types of distortions in group 1.

and for only one distortion type, the SROCC metric is less than 0.85. In group 2, all the six algorithms fail in many cases as these kinds of distortions are not included in the training set. However, VROP still achieves convincing results on some distortion types, *e.g.*, the high frequency noise, multiplicative noise, lossy compression in TID2013, and the LAR in IVC.

In addition, we also conduct non-distortion-specific experiments for each testing dataset as in many previous works [15], [23]. The experiments are performed on the images with all types of distortions in group 1. Table IV shows the experimental results. One can observe that the proposed method achieves promising results on these four datasets, better than the other competitors. Moreover, note that VROP works consistently better than VRAP, which once again verifies the effectiveness of the object oriented pooling strategy experimentally.

Fig. 6 shows the scatter plots of the subjective scores versus the objective scores predicted by BRISQUE, CORNIA and VROP in non-distortion-specific experiments. The plus signs and the black curve represent the test images and the nonlinear mapping function ((11)) respectively. Generally, the objective scores generated by a better approach should correlate more consistently with the nonlinear mapping curve. From Fig. 6, we can see that the objective scores estimated by VROP are more correlated with the subjective ratings than the other two competitors. All the above results support the conclusion that our approach is robust against different datasets.

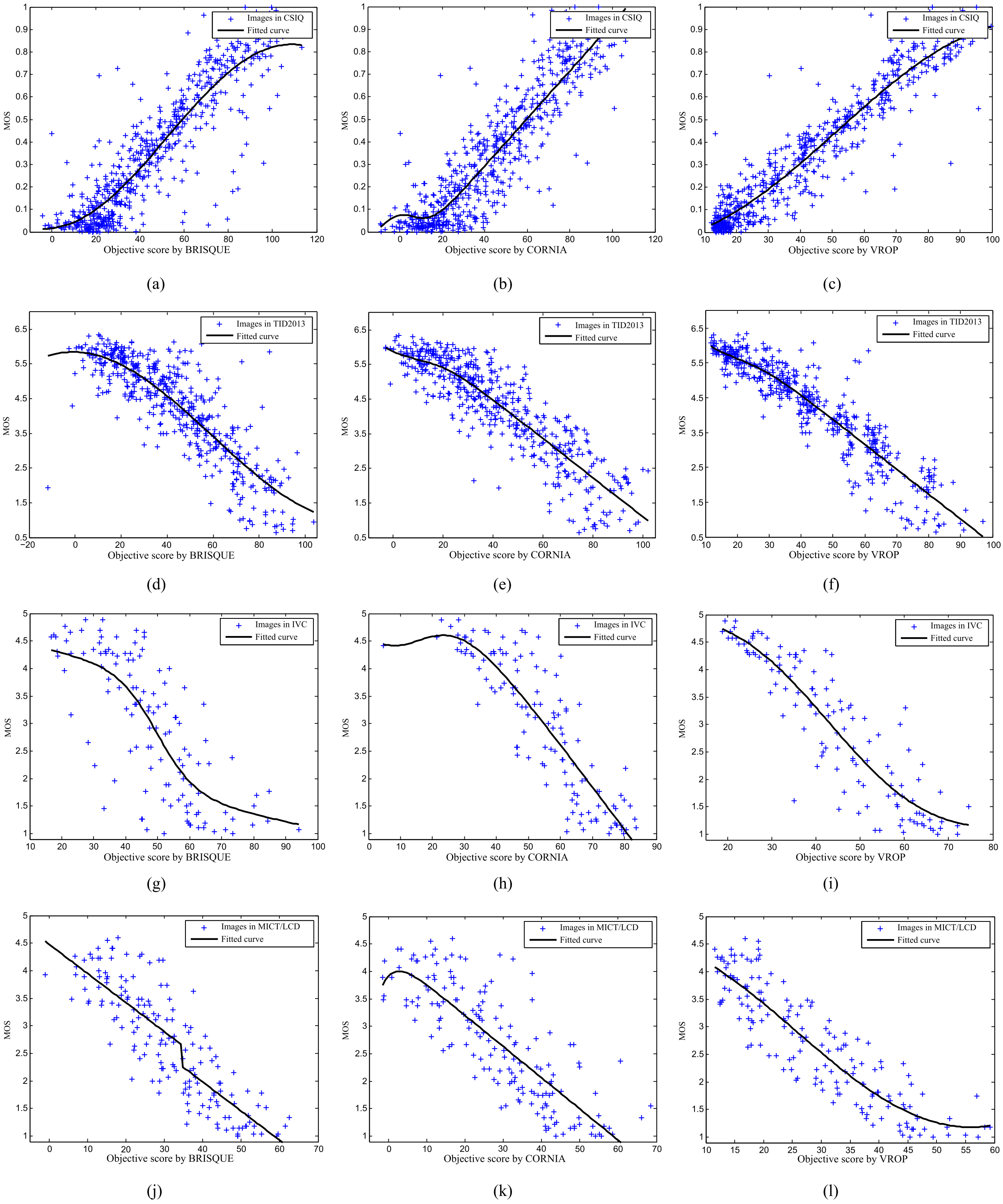
VI. DISCUSSIONS

In this section, we will conduct two extended experiments, compare the single and vector regression framework, and discuss several performance issues.

A. *Evaluation on Images with Authentic Distortions*

Recently, researchers started focusing on IQA with realistic distortions. In this experiment, we evaluate the proposed vector regression framework on two datasets. One is the LIVE In the Wild Image Quality Challenge Database (LIVEW) [61], which contains a total of 1162 images impaired by randomly occurring distortions and genuine capture artifacts. These images, with resolution fixed to 500 × 500 pixels, are captured using a wide variety of modern mobile devices. The other dataset is CID2013 [62]. It contains six image sets, and each set includes six scenes. There are totally 475 images with a relatively large resolution (1600 × 1200).

The main problem in this case is that these authentically distorted images usually contain inhomogeneous impairments,



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 6. Scatter plots of subjective scores versus predicted objective scores on four testing datasets. The models are trained on the LIVE dataset. (a), (d), (g), (j): BRISQUE evaluated on CSIQ, TID2013, IVC, MICT/LCD. (b), (e), (h), (k): CORNIA evaluated on CSIQ, TID2013, IVC, MICT/LCD. (c), (f), (i), (l) VROP evaluated on CSIQ, TID2013, IVC, MICT/LCD.

thus, cropping 32 × 32 image patches for training is no longer appropriate. We address this problem by using networks with larger input size. Specifically, we adapt classification networks (*i.e.*, AlexNet [41], VggNet [30] and ResNet [63]) into the vector regression framework and transfer their learned representations by fine-tuning. Transfer learning is commonly used in many visual tasks to transfer knowledge across different domains [64], [65].

We fine-tune the AlexNet and VggNet (16 layers) by substituting the softmax layer with a 5-D regression layer and then fine-tuning the last two layers to fit the belief scores. ResNet (50 layers) contains less parameters, thus we fine-tune the whole pipeline. The input patch size is fixed to 224 × 224. The iterative optimization is performed about 10 epochs. The learning rate is initially set as 0.001, and then decreased by 0.1 at regular intervals (twice for ResNet, three times for AlexNet and VggNet).

TABLE V
PERFORMANCE COMPARISON ON AUTHENTIC DISTORTIONS

Dataset	LIVEW		CID2013	
Metric	SROCC	LCC	SROCC	LCC
BRISQUE [14]	0.581	0.603	0.668	0.669
CORNIA [15]	0.622	0.653	0.681	0.623
CNN [16]	0.576	0.566	0.526	0.523
NFERM [23]	0.586	0.609	0.600	0.656
SOM [28]	0.624	0.647	0.675	0.622
FRIQUEE [24]	0.694	0.712	0.546	0.552
VR-Alex	0.800	0.818	0.722	0.733
VR-Vgg	0.804	0.835	0.772	0.776
VR-Res	**0.849**	**0.865**	**0.808**	**0.804**

(CNN and SOM are implemented by ourselves.) The networks (namely AlexNet, VggNet and ResNet) integrated with the vector regression framework are abbreviated to VR-Alex, VR-Vgg and VR-Res respectively.

During testing, we average the predicted scores of 50 randomly selected patches to obtain the image level quality score.

For the LIVEW dataset, we randomly divide the dataset into two parts: 80% for training and 20% for testing. The results are reported as the medians across 20 train-test iterations. For the CID2013 dataset, only the image sets IV-VI are used for performance comparison as the other three sets (I-III) are rated with different subjective evaluation protocols. The image sets IV-VI contain a total of eight scenes. Generally, images from the same scene are similar to each other, while different from those from other scenes. Thus we train our model on seven scenes and then test it on the remaining one. This procedure is repeated until all the images have been assigned with the objective scores.

Table V shows the experimental results, where the three networks integrated with the vector regression framework are abbreviated to VR-(Alex, Vgg and Res) respectively. We can see that all these networks can be used as vector regression, and they outperform the current state-of-the-art methods.

B. Single Regression vs Vector Regression

In this section, we aim to demonstrate that the performance of our approach benefits from the vector regression framework. To this end, we use the model without vector regression as the baseline. Specifically, the baseline has the same network architecture as the proposed model, but adopts the single regression framework that maps the extracted features directly to a quality score.

The experiments in Section V-C (Evaluations on TID2008), V-D (Cross Dataset Evaluation) and VI-A (Evaluations on LIVEW) are used to examine the performance. The configurations of the baseline are the same as those of the proposed model, including the experimental procedures and the training configurations in the optimization process. The object oriented pooling is not incorporated as we want to verify the effectiveness of the vector regression.

Table VI, VII and VIII report the experimental results. It can be seen that the networks with vector regression work consistently better than those with single regression. The performance decreases if the vector regression is not used (even the network architecture is the same). The results demonstrate the effectiveness of the vector regression framework. Moreover, we can observe from Table V and VIII that the vector regression can be integrated with different networks and that it can help to improve the performance regardless of the network architecture.

TABLE VI
PERFORMANCE COMPARISON ON THE TID2008 DATASET

Model	baseline		VRAP	
Metric	SROCC	LCC	SROCC	LCC
TID2008	0.840	0.825	0.900	0.893

VRAP is the proposed method that adopts the vector regression framework. The baseline has the same network architecture and training configurations as VRAP, while adopts the single regression framework.

TABLE VII
CROSS DATASET EVALUATION

Model	baseline		VRAP	
Metric	SROCC	LCC	SROCC	LCC
CSIQ	0.911	0.928	0.925	0.936
TID2013	0.888	0.892	0.905	0.911
IVC	0.808	0.811	0.893	0.889
MICT/LCD	0.785	0.777	0.858	0.861

The models are trained on LIVE and then tested on the following four datasets. Only those types of distortions that also appear in LIVE are included. Refer to Table VI for the notations of baseline and VRAP.

TABLE VIII
PERFORMANCE COMPARISON ON THE LIVE IN THE WILD IMAGE QUALITY CHALLENGE DATABASE

Metric	SROCC	LCC
base-Alex	0.784	0.795
VR-Alex	0.800	0.818
base-Vgg	0.789	0.810
VR-Vgg	0.804	0.835
base-Res	0.818	0.837
VR-Res	0.849	0.865

VR-(Alex, Vgg, Res) adopt the proposed vector regression framework, and base-(Alex, Vgg, Res) are the baselines that adopt the single regression framework. The network architectures and training configurations of base-Alex, base-Vgg and base-Res are the same as those of VR-Alex, VR-Vgg and VR-Res, respectively.

C. Discussion on Performance Issues

In this section, we discuss several performance issues about our method, including its sensitivity to the model parameters, the FR measure for selecting training patches, and the sampling scheme in the pooling strategy.

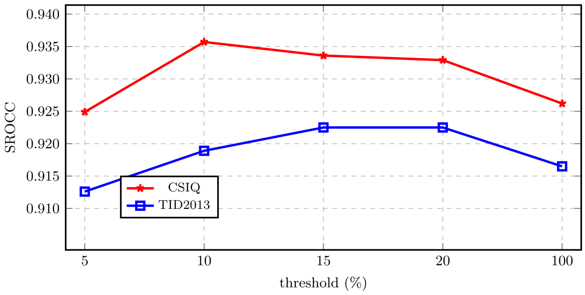


Fig. 7. SROCC with respect to the threshold δ. The models are trained on LIVE and then tested on CSIQ and TID2013.

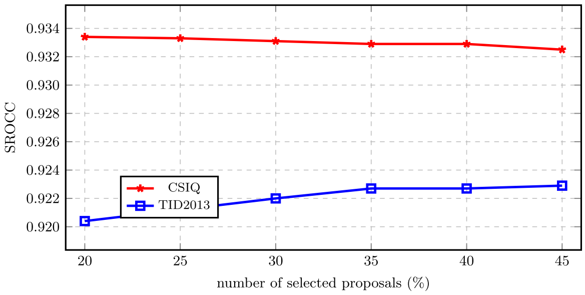
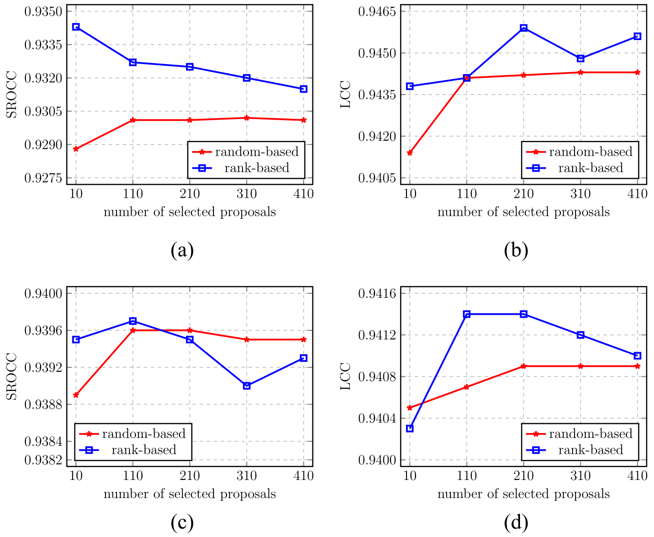


Fig. 8. SROCC with respect to the number of the selected proposals. The models are trained on LIVE and then tested on CSIQ and TID2013.

TABLE IX
VROP-s AND VROP-m REPRESENT THE SINGLE AND MULTIPLE FR BASED MODELS

Model	VROP-s		VROP-m	
Metric	SROCC	LCC	SROCC	LCC
CSIQ	0.933	0.945	0.941	0.946
TID2013	0.923	0.927	0.917	0.918
IVC	0.903	0.901	0.889	0.883
MICT/LCD	0.896	0.898	0.895	0.897

The models are trained on LIVE and then tested on the following four datasets. Only those types of distortions appearing in LIVE are included.



(a) (b)

(c) (d)

Fig. 9. Comparisons between the random-based (red line) and rank-based (blue line) pooling strategies. (a) (b): evaluated on CSIQ. (c) (d): evaluated on TID2013. The horizontal axis represents the number of selected proposals. The rank-based pooling strategy selects top k proposals to guide the pooling, and the other one selects random proposals.

1) Sensitivity to Model Parameters: There are two main parameters in our method, namely the threshold δ for excluding training patches and the number of the selected proposals, *i.e.*, M, for object oriented pooling. In order to evaluate the generalization ability as well, all the results are obtained by training on LIVE and testing on CSIQ and TID2013 (only on those types of distortions appearing in the LIVE dataset).

Threshold for excluding training patches: Fig. 7 shows the SROCC metrics on CSIQ and TID2013 with δ set as 5%, 10%, 15%, 20% and 100% of the data range in LIVE. It is worth noting that none of the cropped patches are excluded in the last case. We can see from Fig. 7 that the performance is relatively poor on both CSIQ and TID2013 when δ is small. This is because too many patches are excluded, and the remaining ones are insufficient to train the deep network. On the other hand, if δ is too large, only few patches are excluded, which will cause a decrease in performance too.

Number of proposals for global pooling: In the global pooling stage, we randomly select a subset of the generated object proposals for object oriented pooling. Fig. 8 shows the SROCC metrics on the CSIQ and TID2013 datasets with different subset sizes. The size of the random subset, namely the number of the selected proposals, is set to be 20%, 25%, 30%, 35%, 40% and 45% of the total number of the generated proposals. We can see from Fig. 8 that the subset size has a minor impact on the estimation results, and the use of more proposals cannot lead to a better performance.

2) Fusion of Multiple FR Metrics: In this part, a more robust way is explored to select image patches in the phase of training data generation. Specifically, we combine multiple FR metrics for a more accurate estimation on the local quality. The method in [18] is employed for the combination. Four FR measures, including GMSD [7], FSIM [5], FSIMC [5] and VSI [8], are used to generate the synthetic scores. The parameters of the combination are set as the default values in [18]. We train the models on LIVE and then compare the performance on other four datasets, namely CSIQ, TID2013, IVC and MICT/LCD.

Table IX lists the results, where the single and multiple FR based models are abbreviated to VROP-s and VROP-m, respectively. One can see that VROP-s and VROP-m have similar performance. Generally, the fusion of multiple FR measures may slightly benefit the local quality estimation. However, it has a minor impact on the training data selection as we only exclude noisy training patches. A patch is excluded from the training set only when the objective score estimated by the FR measure is very different from the ground-truth score of its source image.

3) Rank-Based Pooling Strategy: In object oriented pooling, we randomly select a part of the object proposals to estimate the global quality because the GOP method [49] does not provide scored (or sorted) object candidates. In this part, we investigate

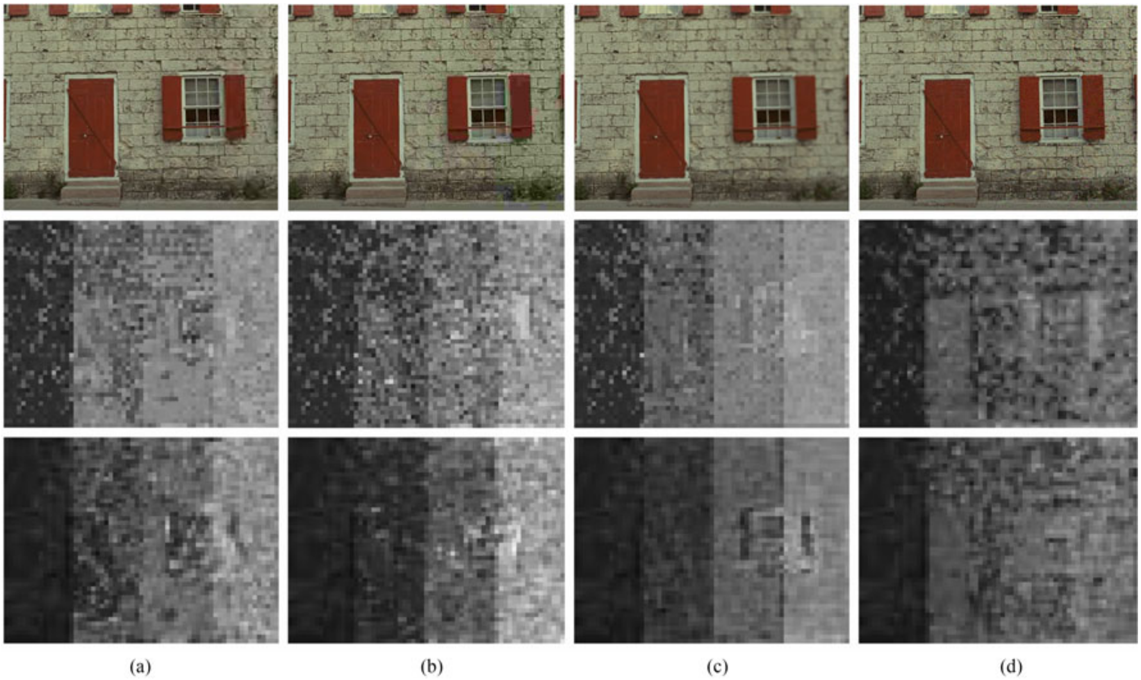


Fig. 10. Illustration of the local quality map, where brighter pixel represents lower quality. The first row exhibits the composite images generated by replacing three vertical areas with distorted versions at different degradation levels. The corresponding distortion types are (a) JP2K, (b) JPEG, (c) BLUR, and (d) WN. The second row shows the local quality maps in [16]. The third row shows the local quality maps generated by the proposed method.

a rank-based object oriented pooling strategy that only selects the top-ranked candidates to guide the pooling. The rank-based pooling may select less object irrelevant candidates compared with the random pooling strategy.

To this end, a generic proposal evaluator (*e.g.*, [66]) is required to assign a rank score to each object proposal. However, scoring each candidate individually is time-consuming in our case. Thus we implement the rank-based pooling by using a window scoring proposal method MCG [48]. Fig. 9 shows the comparisons, where the red and blue lines represent the random-based and rank-based pooling strategies, respectively. We trained the model on LIVE and then tested it on CSIQ and TID2013. It can be seen that the two pooling methods have a comparable performance. The rank-based pooling works slightly better than the other one, especially when the number of selected proposals is small.

D. Local Quality Estimation

The proposed method can naturally give a local quality map (LQM) of the input image by performing pixel-wise averaging on $\mathbf{L}_p$ and $\mathbf{L}_{p+1}$ in (7).

As previously done in [16], an intuitive example is shown in Fig. 10. Our model is trained on LIVE and tested on four synthetic images. We select a reference image from TID2008 and divide it into four vertical parts. Then, for each specific distortion type, including JP2K, JPEG, WN and BLUR, we replace the second to the fourth vertical parts with the distorted versions from high quality to low quality. In this way, a total of four testing synthetic images are generated. We then scan 16×16 patches with a stride of 8 on these synthetic images to obtain their respective LQMs.

Fig. 10 shows the generated LQMs, which have been normalized to $[0, 255]$ for visualization. We can see that the four vertical parts of each composite image are distinguished properly. Compared to the LQMs in [16], our approach generates less noise points and makes a clearer distinction between the adjacent vertical parts, especially on JPEG and BLUR.

E. Issue of Overfitting

We adopted several techniques to prevent overfitting, *e.g.*, dropout, regularization and data augmentation. With these techniques, our model shows a good generalization ability on existing public IQA datasets. In our experiments, we did not observe significant overfitting.

However, the risk of overfitting may still exist in applications because of the limited image contents in current IQA datasets. The training images are generally not sufficient to represent the population of natural images. Some recent studies have discussed the potential overfitting risk when applying current learning based IQA models to real-world suitations [67]. In the future work, unsupervised or semi-supervised network training or feature learning methods could be explored to help reducing the overfitting risk.

VII. Conclusion

In this paper, we have proposed a vector regression framework for BIQA by combining two processes: belief score estimation and object oriented pooling. Specifically, we predict the objective score of an image by first estimating a vector of belief scores. The belief score estimation is implemented by a convolutional network in practice. The object oriented pooling strategy further boosts the performance by incorporating semantic information of image contents. Our approach has shown a great performance and generalization ability in comparison with current state-of-the-art BIQA methods. We also have demonstrated that this framework can be integrated with a pre-trained network and outperforms other competitors on authentically distorted images.

References

[1] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[3] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[4] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.

[5] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[6] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.

[7] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[8] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014.

[9] L. Ma, S. Li, F. Zhang, and K. N. Ngan, "Reduced-reference image quality assessment using reorganized DCT-based image representation," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 824–829, Aug. 2011.

[10] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3378–3389, Aug. 2012.

[11] S. A. Golestaneh and L. J. Karam, "Reduced-reference quality assessment based on the entropy of DWT coefficients of locally weighted gradient magnitudes," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5293–5303, Nov. 2016.

[12] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[13] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[14] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[15] P. Ye, J. Kumar, L. Kang, and D. S. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2012, pp. 1098–1105.

[16] L. Kang, P. Ye, Y. Li, and D. S. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2014, pp. 1733–1740.

[17] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2013, pp. 995–1002.

[18] P. Ye, J. Kumar, and D. S. Doermann, "Beyond human opinion scores: Blind image quality assessment based on synthetic scores," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2014, pp. 4241–4248.

[19] A. Saha and Q. J. Wu, "Utilizing image scales towards totally training free blind image quality assessment," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1879–1892, Jun. 2015.

[20] P. Gastaldo, R. Zunino, and J. Redi, "Supporting visual quality assessment with machine learning," *EURASIP J. Image Video Process.*, vol. 2013, no. 1, pp. 1–15, 2013.

[21] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[22] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2011, pp. 305–312.

[23] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.

[24] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, pp. 32, 2017, doi: 10.1167/17.1.32.

[25] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[26] P. Ye and D. S. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3129–3138, Jul. 2012.

[27] L. He, D. Tao, X. Li, and X. Gao, "Sparse representation for blind image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2012, pp. 1146–1153.

[28] P. Zhang, W. Zhou, L. Wu, and H. Li, "SOM: Semantic obviousness metric for image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 2394–2402.

[29] Q. Wu *et al.*, "Blind image quality assessment based on multichannel feature fusion and label transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 425–440, Mar. 2016.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[31] R. B. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[32] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[33] H. Li, Y. Li, and F. Porikli, "Deeptrack: Learning discriminative feature representations online for robust visual tracking," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1834–1848, Apr. 2016.

[34] D. Ghadiyaram and A. C. Bovik, "Blind image quality assessment on real distorted images using deep belief nets," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2014, pp. 946–950.

[35] H. Tang, N. Joshi, and A. Kapoor, "Blind image quality assessment using semi-supervised rectifier networks," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2014, pp. 2877–2884.

[36] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, Jun. 2015.

[37] W. Hou and X. Gao, "Saliency-guided deep framework for image quality assessment," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 46–55, Apr.–Jun. 2015.

[38] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 990–998.

[39] L. Ma, L. Xu, Y. Zhang, Y. Yan, and K. N. Ngan, "No-reference retargeted image quality assessment based on pairwise rank learning," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2228–2237, Nov. 2016.

[40] T. Judd, K. A. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 2106–2113.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[42] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2014, pp. 580–587.

[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 3431–3440.

[44] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 809–817.

[45] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.

[46] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[47] M. Cheng, Z. Zhang, W. Lin, and P. H. S. Torr, “BING: Binarized normed gradients for objectness estimation at 300fps,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2014, pp. 3286–3293.

[48] P. A. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marqués, and J. Malik, “Multiscale combinatorial grouping,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2014, pp. 328–335.

[49] P. Krähenbühl and V. Koltun, “Geodesic object proposals,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 725–739.

[50] P. G. Engeldrum, *Psychometric Scaling: A Toolkit for Imaging Systems Development*. Winchester, MA, USA: Imcotek Press, 2000.

[51] V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[52] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[53] J. Redi, H. Liu, R. Zunino, and I. Heynderickx, “Interactions of visual attention and quality perception,” in *Proc. SPIE*, vol. 7865, 2011, Art. no. 78650S.

[54] U. Engelke, H. Kaprykowsky, H. Zepernick, and P. Ndjiki-Nya, “Visual attention in quality assessment,” *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 50–59, Nov. 2011.

[55] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[56] N. Ponomarenko *et al.*, “—A database for evaluation of full-reference visual quality assessment metrics,” *Adv. Modern Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[57] N. N. Ponomarenko *et al.*, “Image database TID2013: Peculiarities, results and perspectives,” *Signal Process.: Image Commun.*, vol. 30, pp. 57–77, 2015.

[58] E. C. Larson and D. M. Chandler, “Most apparent distortion: Full-reference image quality assessment and the role of strategy,” *J. Electron. Imag.*, vol. 19, no. 1, 2010, Art. no. 011006.

[59] P. Le Callet and F. Autrusseau, “Subjective quality assessment irccyn/ivc database,” 2005. [Online]. Available: http://www.irccyn.ec-nantes.fr/ivcdb/

[60] S. Tourancheau, F. Autrusseau, Z. P. Sazzad, and Y. Horita, “Impact of subjective dataset on the performance of image quality metrics,” in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 365–368.

[61] D. Ghadiyaram and A. C. Bovik, “Massive online crowdsourced study of subjective and objective picture quality,” *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[62] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen, “CID2013: A database for evaluating no-reference image quality assessment algorithms,” *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 390–402, Jan. 2015.

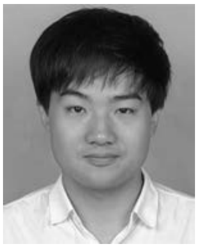
[63] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 770–778.

[64] L. Zhang, W. Zuo, and D. Zhang, “LSDT: Latent sparse domain transfer learning for visual adaptation,” *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1177–1191, Mar. 2016.

[65] L. Zhang and D. Zhang, “Robust visual knowledge transfer via extreme learning machine-based domain adaptation,” *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4959–4973, Oct. 2016.

[66] Q. Wu, H. Li, K. N. Ngan, and L. Xu, “Blind proposal quality assessment via deep objectness representation and local linear regression,” in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2017, pp. 10–14.

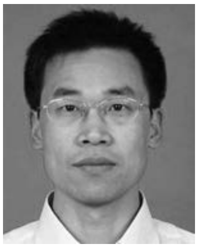
[67] K. Ma *et al.*, “Group MAD competition? A new methodology to compare objective image quality models,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 1664–1673.



Jie Gu received the B.S. degree in applied mathematics from North China Electric Power University, Baoding, China, in 2014. He is currently working toward the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include image processing, computer vision, and machine learning.

Gaofeng Meng (SM’17) received the B.S. degree in applied mathematics from Northwestern Polytechnical University, Xi’an, China, in 2002, and the M.S. degree in applied mathematics from Tianjin University, Tianjin, China, in 2005, and the Ph.D. degree in control science and engineering from Xi’an Jiaotong University, Xi’an, China, in 2009. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include document image processing and computer vision. He is an Associate Editor for *Neurocomputing*.

Judith A. Redi is a Senior Data Scientist at Exact, the Netherlands. She was earlier an Assistant Professor with the Intelligent Systems Department of Delft University of Technology. Her research interests include the perception of quality of multimedia experiences, which connects visual perception, cognitive science, machine learning, and computer vision. Dr. Redi received the Best ICT PhD thesis award from the University of Genoa in 2010 and an NWO-VENI-Grant for her research on semantics and visual quality in 2012. She has been TPC chair of QoMEX in 2015, Area Chair for ACM Multimedia in 2016, and General Chair for ACM TVX in 2017.



Shiming Xiang (M’13) received the B.S. degree in mathematics from Chongqing Normal University, Chongqing, China, in 1993, the M.S. degree from Chongqing University, Chongqing, in 1996, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2004. From 1996 to 2001, he was a Lecturer with the Huazhong University of Science and Technology, Wuhan, China. He was a Postdoctorate candidate with the Department of Automation, Tsinghua University, Beijing, China, until 2006. He is currently a Professor with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition, machine learning, and computer vision.

Chunhong Pan (M’14) received the B.S. degree in automatic control from Tsinghua University, Beijing, China, in 1987, the M.S. degree from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 1990, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2000. He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, image processing, computer graphics, and remote sensing.